Assignment 3: Mastering RandomForest, XGBoost, t-SNE

W. Kowalczyk

wojtek@liacs.nl

17.11.2020

Introduction

The purpose of this assignment is to master two most popular and powerful algorithms for classification and regression tasks: RandomForest and XGBoost. Additionally, you should master three algorithms for dimensionality reduction and data visualization: PCA, LLE and t-SNE. By "mastering algorithms" we mean here the ability of:

- installing these algorithms on your computers,

- downloading and preprocessing (when needed) relevant datasets,

- applying the algorithms to the data,

- demonstrating results.

You are free to search the internet to find some interesting datasets, examples, blogs, etc. However, in your final submission, you should provide references to these sources. And, needless to say, your work should be substantially different than "copy & paste & get-it-working"!

As a starting point we recommend reading the following papers and visiting the sites:

Random Forests: Chapter 15 of the ESLII book:

https://web.stanford.edu/~hastie/Papers/ESLII.pdf

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

XGBoost:

https://xgboost.readthedocs.io/en/latest/tutorials/model.html

https://arxiv.org/pdf/1603.02754.pdf

https://github.com/dmlc/xgboost/blob/master/demo/guide-python/sklearn_examples.py

t-SNE and others:

https://www.youtube.com/watch?v=RJVL80Gg3lA&list=UUtXKDgv1AVoG88PLl8nGXmw

https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf

https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf

Sources of data:

https://archive.ics.uci.edu/ml/index.php

https://www.kaggle.com/datasets

https://www.kdnuggets.com/datasets/index.html

What to deliver?

Three, self-contained Jupyter notebooks that cover your experiments with RandomForest, XGBoost, Visualization. Additionally, a report (pdf) containing the following sections:

- The problem and the data.
- Experimental setup:
  - Data preprocessing,
  - Exploratory data analysis,
  - Base-line accuracy,
  - Parameter tuning.
- Results.
- Conclusions.

Results of data visualization, depending on the problem/data, can be included either in the Data preprocessing or the Results part.

Evaluation Criteria:

- Quality of the code.
- Quality of the report (standard criteria).
- Reproducibility (the code should run on TA's computers!).

DEADLINE: will be announced soon.